Information and Employee Evaluation:
Evidence from a Randomized Intervention in Public Schools

Jonah E. Rockoff
Columbia Business School

Douglas O. Staiger
Dartmouth College

Thomas J. Kane
Harvard Graduate School of Education

Eric S. Taylor
Stanford University

August, 2011[*]

Abstract

We examine how employers learn about worker productivity in the context of a randomized pilot experiment which provided objective estimates of teacher performance to school principals. We test several hypotheses that provide support for a simple Bayesian learning model with imperfect information. First, the correlation between performance estimates and prior beliefs rises with more precise objective estimates and more precise subjective priors. Second, new information exerts greater influence on posterior beliefs when it is more precise and when priors are less precise. Employer learning also affects job separation and productivity in schools, increasing turnover for teachers with low performance estimates and producing small test score improvements.

[*] Corresponding author: jonah.rockoff@columbia.edu. We thank seminar participants at Northwestern University, Stanford University (Business School and School of Education), Harvard University (Economics Department, Kennedy School, and School of Education), Columbia University (Business School and Teachers College), University of Virginia, University of Florida, UC Davis, University of Oregon, University of Tel Aviv, the Research Institute of Industrial Economics (IFN), and the London School of Economics for many helpful comments and suggestions. Financial support was provided by the Fund for Public Schools. All opinions expressed herein represent those of the authors and not necessarily those of the New York City Department of Education.

A substantial theoretical literature focuses on how employers with imperfect information learn about employee productivity, with seminal work by Spence (1973) and Jovanovic (1979). In contrast, empirical research on employer learning has developed more slowly, particularly with regard to subjective evaluation of employees (see Farber and Gibbons, 1996; Oyer and Schaefer, 2011).[1] In this study, we use micro-data to examine how new information on worker performance is incorporated into employee evaluations. Specifically, we analyze the results of a randomized pilot program where principals in New York City schools received new performance measures on their teachers based on student test score outcomes.

In addition to providing a rigorous and detailed test of how new information affects employers' subjective beliefs about worker productivity, our study is relevant to current debates on the use of student outcomes to measure teacher performance. A series of studies demonstrate that productivity varies greatly across teachers (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007), and two recent papers argue in favor of using estimates of teacher "value-added"—test-based performance measures—to evaluate teachers (Gordon et al., 2006; Kane and Staiger, 2008). Nevertheless, some have questioned whether such measures are sufficiently reliable and valid to be used in personnel decisions (Baker et al., 2010; Corcoran, 2010).

Previous research in economics provides considerable support for the distribution of standardized information on teacher performance to school principals for use in making managerial decisions. Although their decision-making power is often constrained by union contracts and other regulations, there is mounting evidence that principals influence student

[1] The most well-known empirical studies on this topic focus on the consistency of wage trajectories with employer learning (e.g., Murphy, 1986; Gibbons and Katz, 1991; Farber and Gibbons, 1996; Altonji and Pierret, 2001; Lange, 2007). Related work has provided descriptive evidence on how managers form employee evaluations. Baker et al. (1988) and Murphy and Oyer (2001) document variation in the use of subjective and objective evaluations across occupations, while other studies highlight managers' reluctance to differentiate among employees (e.g., Medoff and Abraham, 1980; Murphy, 1992) and potential bias in subjective evaluations (see Prendergast and Topel, 1993).

achievement in their role as organizational and instructional leaders (Grissom and Loeb 2010, Clark et al. 2009). Principals provide guidance and feedback to teachers based on classroom observation, allocate resources for training and professional development, and make the ultimate decisions regarding teaching assignments and employment at the school.

Principals' opinions on teacher performance exhibit considerable variation, and they dismiss low performing teachers when given incentives to do so (Jacob, 2007; Jacob and Lefgren, 2008; Jacob 2010). However, some principals may lack the capacity to analyze student data to make judgments on teacher performance, may have little scope with which to compare their teachers to those working in nearby schools, and may have little training in teacher evaluation. Distributing standardized information on teacher performance based on student achievement to principals can address these limitations, and the verifiability of these measures can help principals act on their beliefs regarding employee performance (Baker, Gibbons, and Murphy, 1994). Furthermore, allowing principals to process this information and make holistic evaluations does not provide strong incentives for dysfunctional behaviors (Holmstrom and Milgrom, 1991).

Our empirical analysis is based on a simple Bayesian learning model in which principals use imperfect information to learn about teacher productivity. This provides us with several empirical predictions for the relationship between teacher performance data and principals' prior and posterior beliefs. All of the model's predictions are borne out in our data. There is a strong relationship between teacher value-added measures and principals' baseline evaluations of teacher effectiveness, and this relationship is stronger when value-added estimates and principals' priors are relatively more precise. More importantly, principals incorporate the objective performance data into their posterior beliefs, and do so to a greater extent when the new information is more precise and when their priors are less precise. Provision of this new information also led to a change in patterns of retention (i.e., higher value-added teachers were more likely to be retained) and small improvements in student achievement the following year.

In Section 2 we present information on the program we study, surveys taken at baseline and follow-up, other sources of data, and the treatment and control groups. In Section 3, we discuss the basic theoretical framework which guides our analysis of the impact of the treatment. We examine the relationship between the value-added performance estimates and principals' prior beliefs regarding teachers in Section 4, and we estimate impacts of the provision of performance data in Sections 5 and 6. Section 7 concludes.

2. Program Timeline, Data Collection, and Treatment

In 2007, the New York City Department of Education (hereafter NYC) enacted a set of policies to give principals more decision-making power and responsibility in improving the performance of their schools. This included a new funding system based on student enrollment, greater flexibility for principals to allocate funds as they saw fit, and the creation a new school accountability system where principals in poor performing schools could lose their jobs and those in successful schools could receive bonuses of up to $25,000 (Rockoff and Turner, 2010).

As part of this policy shift, NYC started a pilot program to provide estimates of teachers' "value-added" (i.e., impacts on student test scores) to principals in order to help them manage their schools.[2] A timeline of the pilot program is shown in Figure 1. On July 10, 2007, just over

[2] NYC officials were aware of academic debates regarding the validity of assumptions underlying value-added (see Todd and Wolpin, 2003, Rothstein, 2009, Harris and Sass, 2006, Goldhaber and Hansen, 2009, Koedel and Betts, 2009, and Staiger and Rockoff 2010), and felt that principals would have local knowledge regarding student assignment that could enable them to interpret value-added estimates in a meaningful way.

1000 principals in schools containing grades 4, 5, 6, 7, or 8 were sent a recruitment e-mail.[3] It contained a description of the initiative, an invitation to information sessions being conducted on July 17, 19, and 23, and a response link to indicate interest in participating. Two weeks later, 305 principals had indicated an interest in participation, meeting NYC's goal of 300 volunteers, and enrollment was closed on July 30 with a total of 335 volunteers.

These principals were sent a baseline survey on August 8, 2007, which they had to complete by September 21 to be eligible to receive value-added reports. The main focus of the survey, which we describe in more detail below, was to solicit principals' performance evaluations of teachers for whom value-added estimates could be made. The 223 principals who completed the survey on time were stratified by grade configuration (elementary, middle, and K-8 schools), and 112 were selected into the treatment group via a random number.

A sample value-added data report is included as Figure 2. A report contained four different value-added estimates in each subject (math and/or English), as well as confidence intervals for each estimate.[4] Each teacher's performance was compared to teachers citywide and to "peer" teachers with similar levels of experience working in classrooms with similar student composition. For both citywide and peer comparisons, value-added was measured based on up to three years of prior data and on just the prior year of data.

[3] Principals were contacted by personnel at The Battelle Memorial Institute, which worked under contract with NYC to recruit principals, perform randomization into treatment and control groups, collect survey data, estimate teacher value-added, prepare the value-added reports, and provide professional development to principals.

[4] We do not discuss the value-added estimation methodology here. Details were provided to participating principals in a technical report which is available upon request from the authors. The methodology uses linear regression to predict student test scores based on prior information, and averages residuals at the teacher level. Because the standardized tests are taken prior to the end of the school year, teachers' value-added estimates are based partially on test score performance of students in the following year. This partial weighting on next year's performance was not done for teachers of 8th grade (due to a lack of 9th grade test data) or for teachers observed for the first time in the school year 2006-2007 (the most recent year of data used to estimate value-added). Value-added measures were based only on the students assigned to teachers at their current schools and were calculated separately by grade level. For teachers who had taught multiple grades, multiple reports were received, and we average value-added across these reports, weighting by the number of students taught. While the reports also present value-added estimates specific to subgroups (e.g., English Language Learners, Special Education students), we restrict our analysis to estimates based on all students.

Treatment principals began receiving these reports in December 2007, along with training on the methodology used to estimate value-added. 71 principals attended a training session and received reports in person in December, 2007. Between mid-January and late-February of 2008, 24 principals participated in remote sessions (i.e., a conference call and on-line presentation) and 1 viewed a session on video. These principals received their reports, but the remaining 16 treatment principals not attending training did not receive any reports.

In late May 2008, a follow-up survey was sent to all principals save two in the treatment group and one in the control group who asked to be removed from the study. Principals were asked to provide another evaluation of each teacher for whom a value-added estimate was made. The survey was closed in mid-July, with an overall response rate of 80 percent.

In addition to the elements of the pilot program, several other events occurring during this period are relevant to the interpretation of our findings. First, while the identities of participating principals were not made public, the pilot's existence was made well known in reports by the popular press (e.g., New York Times, January 21, 2008). The NYC teachers' union did not support the program, and, in April 2008, the city and state unions successfully lobbied a change in state law so that teachers could "not be granted or denied tenure based on student performance data."[5] Treatment principals were advised by NYC that the reports were not to be used for formal teacher evaluation during the pilot year.

Finally, in December of 2008, NYC distributed teacher value-added reports to all principals in the district whose schools served any grade from 4 to 8. Thus, when we examine performance on tests taken in English (Math) in January/March 2009, control principals had received value-added reports, albeit only a short time prior to testing dates.

[5] This law was repealed in 2010.

2.1 Survey Data

First and foremost, the surveys solicited principals' evaluations of their teachers. Principals were asked to evaluate each teacher "overall" and "in terms of raising student achievement" in math, English, or both (for teachers of both subjects). Evaluations were on a six point scale: Exceptional (top 5 percent), Very Good (76-95th percentile), Good (51-75th), Fair (26-50th), Poor (6-25th), or Very Poor (bottom 5 percent).[6]

Descriptive statistics on baseline evaluations are presented at the top of Table 1. On the 1-6 scale, the mean overall evaluation was 4.3, with a standard deviation of 1.1 points. The distribution is skewed, with three quarters of teachers receiving an overall evaluation indicating above median performance (Good, Very Good, or Exceptional).[7] Still, consistent with theory (Macleod 2003), the survey indicates much more variation than do formal teacher evaluations; in NYC, 98 percent of teachers are deemed "Satisfactory" and just 2 percent "Unsatisfactory."[8]

Principals were also asked to provide the number of formal classroom observations and total classroom observations they had made of each teacher during the prior school year.[9] On average, principals reported 2.2 formal observations and 6.4 total observations of each teacher. There was considerable variance across principals in the frequency of observation, with 20

[6] As with the value-added measures, the survey focused on teachers who taught math and/or English to students in grades four through eight. Principals were directed to compare each teacher to all "teachers [they] have known who taught the same grade/subject," not just to teachers within their school or with similar experience.

[7] From Exceptional to Very Poor, percentages in each group were 13.7, 33.6, 30.2, 17.2, 3.8, and 1.6, respectively.

[8] See Weisberg et al. (2009) for more evidence on principals' reluctance to give teachers poor formal evaluations. As one might expect, teachers that received unsatisfactory formal evaluations in the prior school year were given an average rating of just 1.93 on the survey's 1-6 scale, relative to 4.37 for teachers given satisfactory evaluations.

[9] Formal observations are part of the official NYC teacher evaluation system. Untenured teachers must be formally observed at least twice per year. Observations typically last one full lesson, though it is permissible under the union contract to have a series of short observations covering different aspects of a series of lessons. The principal must hold a discussion with the teacher before and after the lesson is observed, hold a post-observation meeting with the teacher, and provide recommendations to improve the teacher's instruction. Tenured teachers must be formally observed once, or evaluated via "performance option" which entails setting goals at the start of the school year and submitting a report to the principal at the end of the year on how those goals were met. In our analysis we code "more than ten" observations as having a value of 11.

percent reporting no more than four total observations of *any* teacher, and 15 percent reporting "more than ten" total observations of *every* teacher in their school.

The baseline survey also asked principals about various management issues relating to teacher evaluation: how they assess teachers, their opinions on measuring teacher performance using student test scores, their ability to attract and retain high quality teachers, etc. Descriptive statistics on these survey items are provided in Appendix Table A1. Of note is that more than three quarters of principals strongly agreed with the statement "I know who the more and less effective teachers are in my school," while less than half the principals agreed with the statement "I am able to dismiss ineffective teachers in my school," and only a quarter agreed with the statement "anyone can learn to be an effective teacher." Thus, the principals had fairly strong priors, were concerned about the presence of ineffective teachers in their schools, and did not believe that training could lead all of their teachers to become effective.

The follow-up survey in May 2008 also solicited evaluations and observation frequency on teachers. In addition, all principals were asked about teacher evaluation and treatment principals were asked about the value-added reports and how they were used. We discuss these items in more detail below.

2.2 Other Data Sources

We use data from teacher value-added reports for all schools in the city, not just from the treatment schools whose principals received them. For ease of interpretation, we normalize value-added measures at the city level to have a mean of zero and standard deviation of one.[10] These reports also provide us with a categorical variable for teaching experience, the number of

[10] In order to be familiar to principals, value-added measures were reported in "proficiency rating units," a scale used by NYC in its accountability system. The variance of value-added estimates prior to normalization was higher for those based on one year of data than for those using up to three, and higher for math than English, consistent with other studies from New York and elsewhere (see Hanushek and Rivkin (2010) for a review).

years the teacher had been teaching a particular grade/subject within their school (up to a maximum of three), and some aggregate information on the demographics and achievement levels of the students whose test scores contributed to the teachers value-added estimates.

Human resources records provide us with information on whether a teacher switched to another school within NYC or left NYC's teacher workforce. They also provide information on the demographics and work experience of principals in our sample, which we use to test the validity of randomization and to measure the number of years a principal has observed teachers in their current capacity.

Finally, we have data on students' test scores, demographics (i.e., gender, race/ethnicity, free lunch receipt, English language learner, and special education status), and links to the students' math and English teachers for the school year 2008-2009. Due to turnover and changes in teaching assignments, only 58 percent of teachers in our baseline sample were teaching math or English to students in grades 4 to 8 in the same school in the school year 2008-2009.

2.3 Randomization and Sample Selection

To check that the randomization was successful, we compare the average characteristics of treatment and control principals and schools, including principals' work experience and demographics (from HR records), students' characteristics (from enrollment data), and teachers' opinions on school management issues, which were collected in a survey unrelated to our study in the spring of 2007. These comparisons (shown in the leftmost columns of Table 2), show the two groups were very similar, and in no case do we find any statistically significant differences.

Limiting our sample to principals that responded to the follow-up survey, we find no significant differences between treatment and control characteristics at baseline (Table 2, middle columns), supporting the notion that the groups were also comparable at follow-up.[11] For our analyses of non-survey outcomes (e.g., teacher turnover, student test scores), we include all participating schools, regardless of response to the follow-up survey.

Table 1 also provides descriptive statistics for teacher level variables used in our analysis. Again, we find treatment and control groups are quite similar. There is a marginally significant difference in exit rates between when our baseline survey was sent out and start of the school year (16 vs. 13 percent). This was prior to randomization and, given the number of variables we have tested, is likely due to chance, but we examine "placebo" effects of value-added information on turnover during this pre-experimental period as a robustness check.

While we cannot rule out the existence of unobservable factors which caused some principals to sign up for the pilot program, we find that participating principals (schools) were similar to the NYC population on a wide range of observable characteristics (Table 2, rightmost columns). We find only a few statistically significant differences that are small in magnitude, e.g., sample principals had more teaching experience (7.2 vs. 6.2 years) and were more likely to be female (79 vs. 73 percent). More importantly, we find no significant differences on teachers' opinions with regard to issues such as the frequency with which the principal conducts classroom observations, the quality of feedback teachers receive from the principal, how much the principal prioritizes teaching, or the use of student data in instruction. In addition, the take-up rate for the pilot program—about one in five—does not seem particularly low, considering that recruitment was based on a single e-mail during the summer and enrollment was closed after less than three weeks.[12] This suggests our findings regarding employer learning are applicable to the population

[11] Response rates at follow-up were lower for treatment schools (75 percent) than control (85 percent). This was mainly due to the 16 treatment principals who never received reports and had a response rate of 30 percent.

[12] This level of take-up is also on par with other studies. Take-up for Project STAR, a well-known class size reduction experiment (see Krueger, 1999; Chetty et al., 2010) that provided schools with additional teachers was

of NYC principals, to school leaders in other districts around the country, and likely to other managers evaluating performance in skilled occupations where performance is multi-faceted.

2.4 Face Validity of Treatment

The creation of new performance reports does not automatically imply that managers will read and understand them, let alone use them in making strategic decisions. In an examination of new school performance measures in Pennsylvania, McCaffrey and Hamilton (2007) find that nearly half of school principals either had no knowledge of the new system or claimed never to have seen any of the new reports. Here we discuss evidence that most, though not all, of the principals in the treatment group received, examined, and understood the value-added reports, and that selection into this group has "face validity" as a measure of the intended treatment.

As mentioned above, 16 treatment principals were not given reports on their teachers because they did not attend training around value-added measurement.[13] An additional 13 treatment principals reported on the follow-up survey that they did not examine the value-added reports, which provides an upper bound of 75 percent on the fraction of treatment principals that might have made use of this information. However, principals that claimed to have examined the reports provided largely correct answers to questions regarding the inclusion of various control variables in the measurement of value-added, and 85 percent of them rated the reports as either "Very Useful" or "Useful" (the top two possible ratings) on a five-point scale.[14] We

roughly one in five. Take-up in a recent field experiment that offered free management consulting to mid-sized Indian manufacturing firms (Bloom et al., 2010) was roughly one in four.

[13] Because training and the reports are linked, we cannot determine effects of providing one or the other in isolation. Our approach is based on the view that information is the treatment, but training facilitates use of the information.

[14] Most treatment principals expressed confidence that the value added methodology accounted for actual control variables (e.g., teaching experience, prior test scores, class size) and far fewer thought it accounted for factors not included as controls (e.g., the presence of classroom aides or whether a teacher's students received outside help). See Appendix Table A2 for descriptive statistics on these responses to the follow-up survey.

therefore conclude that treatment status in the pilot is a good but imperfect indicator for whether a principal received, read and understood the information in the teacher value-added reports.

3. Conceptual Framework and Empirical Predictions

We use a simple Bayesian learning model to consider the principal's evaluation problem. Principals accumulate information regarding the performance of each teacher and use this information to construct a belief regarding the teacher's effectiveness. Following earlier work (Jovanovic 1979, Harris and Holmstrom 1982, Gibbons et al. 2005), we assume beliefs are normally distributed (see Equation 1), with expected value μ_0 and precision h_0. This provides us with simple mathematical expressions for the impact of new information.

$$(1)\quad \mu \sim N\left(\mu_0, \frac{1}{h_0}\right)$$

Over time, principals gather new information through various channels such as classroom observation. Equation 2 provides expressions for information routinely accumulated over time (ε) and the information provided in the value-added estimates (V), both of which are assumed to be imperfect signals of teacher effectiveness.[15]

$$(2)\quad \varepsilon \sim N\left(\mu, \frac{1}{h_\varepsilon}\right) \quad V \sim N\left(\mu, \frac{1}{h_V}\right)$$

Following DeGroot (1970), the principal's posterior expectation of teacher effectiveness at time t takes on a simple expression, a weighted average of the expected value of the prior

[15] Note we have assumed that the error components of the value-added estimate to be independent of the principal's prior belief. Relaxing this assumption and allowing for a positive correlation of these error components reduces the extent to which value-added provides new information but does not affect the qualitative implications of the model. In addition, our framework considers teacher effectiveness as a single dimension, though principals (and society) are likely to value multiple dimensions of teaching, some of which may be unrelated to value-added. Our empirical predictions would be unaffected by an alternate framework in which (1) we model beliefs regarding just one dimension of teaching and (2) principals' overall evaluations are a weighted average of their beliefs over all dimensions of teaching, with persistent weights based on principal's preferences.

belief (μ_0) and the information (ε, and possibly V) acquired during the period from t-1 to t. The weights are based on the relative precision of each component (see Equation 3).

$$(3)\ \ E(\mu \mid V,t) = u_1 = \begin{cases} \text{If treatment}: \dfrac{h_0\mu_0 + h_\varepsilon\varepsilon + h_V V}{h_0 + h_\varepsilon + h_V} \\ \text{If control}: \dfrac{h_0\mu_0 + h_\varepsilon\varepsilon}{h_0 + h_\varepsilon} \end{cases}$$

It is straightforward to show this simple model yields two intuitive empirical predictions:

Prediction 1: Value-added estimates (V) should be correlated with principals' prior beliefs (μ_0), since both are signals of teacher effectiveness. This correlation should be stronger when value-added estimates and prior beliefs are more precise (i.e., greater values of h_V and h_0).

Prediction 2: Conditional on principals' prior beliefs (μ_0), treatment principals' posterior beliefs regarding teacher effectiveness should, relative to control principals, place more weight on value-added estimates and less weight on prior beliefs. Treatment principals should place more weight on value-added estimates with greater precision (h_V), and less weight on value-added estimates when their priors are more precise. Principals' posterior beliefs in the control group may also be conditionally correlated with value-added estimates, to the extent that value added is correlated with the new information gathered between surveys by all principals.

Empirically, we use principals' baseline evaluations to measure prior beliefs (μ_0), principals' follow-up evaluations to measure posterior beliefs (μ_1), estimates from the value-added reports to measure V, and the confidence intervals given in the value-added reports to measure h_V. We do not have a readily available measure of the precision of principals' prior beliefs (h_0), but we proxy for it using data on the length of the working relationship between the principal and teacher.[16]

While we focus primarily on how new information affects principals' beliefs, we also test several hypotheses about how it affects principals' actions. First, if classroom observation is costly, then principals could respond to the provision of value-added data by observing teachers

[16] We lack data on teachers' complete work histories but we know their years worked in value-added subjects/grades in the current school. To measure years during which the principal had likely observed the teacher, we take the minimum of this variable and the number of years the principal worked in the current school. We do know teachers' total years of experience, but this is likely to overestimate experience within a school for a large fraction of teachers.

less frequently; the marginal benefit of an additional classroom observation on the principal's posterior belief will be smaller when the principal has more precise information from other sources. Second, it is reasonable to believe that providing value-added information may create a stronger relationship between value-added estimates and teacher turnover, either through changes in principals' posterior beliefs, or providing independent and verifiable confirmation of their priors, along the lines of Baker et al. (1994).[17] Finally, principals may use the information to improve educational quality, in which case we would expect to see increased student achievement relative to control schools.

4. Value-added Estimates and Principals' Priors

The first prediction from our framework—that value-added estimates and principals' prior beliefs should be positively correlated—has been shown before (e.g., Murnane, 1975; Armor et al., 1976; Jacob and Lefgren, 2008; Harris and Sass, 2009), and we confirm this finding. Documenting how precision mediates the strength of this relationship is unique to our study, along with several other results discussed below.

4.1 Baseline Correlations of Value-Added with Prior Beliefs

We document the relationship between value-added estimates and principals' prior beliefs using linear regression, where the evaluation *R* given to teacher *i* at baseline (*t-1*) is regressed on a teacher's value-added estimate ($V_{i,t-1}$), as shown in Equation 4.

$$(4)\ \ R_{i,t-1} = \alpha + \beta V_{i,t-1} + \xi_{i,t-1}$$

In most specifications we pool across math and English, taking the average value-added estimate across both subjects. Principal evaluations and value-added estimates are normalized to have a

[17] Existing studies find that principals are reluctant to remove teachers for poor performance, even when the administrative costs of doing so are minimized (see Jacob, 2007, 2010), suggesting reputational or social costs associated with dismissing a teacher are important.

mean of zero and standard deviation of one, and standard errors are clustered by school. We pool treatment and control groups for power, but our findings are not substantially or statistically different if we examine each group separately.

As expected, there is a strong positive relationship between principals' evaluations of overall performance and value-added estimates. In Table 3, Columns 1 to 3, we find a similar effect size of value added (0.21 to 0.23) for estimates based on (1) multiple years of data and peer teacher comparisons (i.e., those with similar experience and similar classrooms of students), (2) one year of data and peer comparisons, or (3) multiple years of data and citywide comparisons. [18] However, when we run "horse races" between these different value-added estimates (Columns 4 to 6), we find that multi-year estimates based on peer performance are the most robust predictors of principal evaluations.[19] In the remainder of our analysis, we measure value-added using the multi-year peer comparison.

Previous research has found that the relationship between value-added estimates and subjective beliefs regarding teacher effectiveness is somewhat stronger in math than English (Jacob and Lefgren, 2008; Rockoff and Speroni, 2010). However, when we estimate regressions separately by teachers of math, English, or both math and English, we find that principals' prior

[18] These estimates are also robust to controlling for principal fixed effects, and we find very similar results using the principal's evaluations of the teacher's ability to raise student achievement in math and/or English as the dependent variable. This similarity is not surprising given their correlation of the overall evaluation with subject specific evaluations (0.87 for math and 0.88 for English).

[19] Though multi-year estimates using the peer comparison and those based on citywide comparisons are both statistically significant when included together in the regression (Column 5), the conditional relationship between citywide value-added and principals' priors disappears if we control for teacher experience (Column 6), as one might expect. Coefficient estimates on experience indicators are not reported but, as might be expected, baseline evaluations were lowest for teachers who just completed their first year and tend to rise with experience.

beliefs are not clearly tied more strongly to one of the two subject areas, and we always estimate positive and significant coefficients on value-added (see Appendix Table A3).[20]

4.2 Baseline Relationships and the Precision of Information

Moving to the more novel element of Prediction 1, we test whether increased precision of value-added and principals' priors strengthen the relationship between the two variables. To test this prediction, we estimate specifications that include interactions of value-added or the principal's baseline evaluation with a measure of precision (*h*), as shown by Equations 5 and 6:

$$(5)\ R_{i,t-1} = \alpha + \beta\left(1 + \delta^{V} h_{V}\right)V_{i,t-1} + \xi_{i,t-1}$$

$$(6)\ V_{i,t-1} = \tilde{\alpha} + \tilde{\beta}\left(1 + \delta^{R} h_{R}\right)R_{i,t-1} + \tilde{\xi}_{i,t-1}$$

Estimating Equation 5, we test the hypothesis that the coefficient on value-added increases when the value-added estimate is estimated more precisely (i.e., δ^{V} is positive). If value-added is a noisy measure of true teacher performance, then getting a more precise measure should mitigate attenuation bias from classical measurement error and increase the coefficient on value-added. In Equation 6, we simply switch the principal's prior and value-added as dependent and independent variables; this allows us to perform a similar test on whether increased precision in the principal's prior also reduces attenuation bias.

To measure the precision of value-added (h_V) we use the inverse of the confidence interval provided to principals on the value-added report.[21] Consistent with Prediction 1 of our framework, value-added estimates estimated with greater precision have greater predictive power

[20] For teachers of only math, the coefficient on value-added is 0.4, while for teachers of only English the effect size is 0.2. However, among teachers of both subjects, the coefficient estimates were larger for value-added in English than in math, either when estimated separately or together in the same regression.

[21] We find very similar results to those described here if we use an interaction with the inverse of the variance of the value-added estimate, which is a more proper measure of precision. However, we use the inverse of the confidence interval later in our analysis to test Prediction 2, because the confidence interval is what was actually provided to principals in the value-added report. We therefore also use it here for consistency.

for principals' priors (Table 4, Column 1). This result is robust to the inclusion of controls for teacher experience, school fixed effects, and the number of years of data used in generating the value added measure (Columns 2 and 3).[22] We also find positive and significant interactions of value-added and precision when separately examining teachers by subject area.

As mentioned in Section 3, we lack a direct measure for the precision of the principals' prior beliefs. We therefore assume that beliefs become more precise over time, and proxy for precision using the length of the working relationship between principals and teachers. We limit the sample to teachers we are sure have worked in the school for at least three years, so that more experienced principals' evaluations are sure to be based on more information than those made by principals with less experience, and test for a positive coefficient on the interaction of baseline evaluation with the length of time the principal has been at the school (h_R).

The empirical results are in line with this expectation and the Bayesian model. Both the main effect of principals' evaluations and the interaction with principal experience are positive and (marginally) statistically significant (Table 5, Column 1).[23] One potential issue with this specification is that experienced principals are likely to have information that is not captured by

[22] The motivation behind these controls is that teachers for whom more years of data are available to generate value-added estimates may tend to be both more effective, even conditional on total teaching experience, and have smaller confidence intervals. While we do not report coefficients on the main effect of precision in Table 4, this is included in all of the regressions and its estimate is positive and statistically significant. We find that this is driven by principals giving higher ratings to teachers working with low scoring students (i.e., those in the bottom tercile of the city distribution); adding a control for the fraction of the teachers' students scoring in the bottom tercile—on which the coefficient is positive and highly significant—causes the main effect of precision to fall from 0.103 (t-stat 2.8) to 0.024 (t-stat now 0.8), but has little impact on the main effect of value-added or the interaction of value-added with precision, both of which remain highly significant. Adding an interaction of the fraction of students in the bottom tercile with the precision of value added—on which the coefficient is small and insignificant—also has little impact on the results in Table 4.

[23] Although not reported in Table 5, we include a main effect of our precision measure and find its coefficient is always small and statistically insignificant. If we use evaluations of teachers' ability to raise student achievement instead of the "overall" evaluation as the dependent variable, we get slightly larger point estimates on the main effect and interaction (0.158 and 0.015, respectively), and both are significant at the 5 percent level.

the available value-added estimate, which is only based on three years of data.[24] We therefore would expect to find a larger coefficient on the interaction of the principal's evaluation and principal experience when we limit the sample to teachers with fewer years of experience. This is indeed the case, with interaction terms growing as we remove teachers with 10 or more years of experience (Column 2) and teachers with 5 or more years of experience (Column 3).[25]

This analysis confirms both elements of Prediction 1 regarding the baseline relationship between value-added measures and principal evaluations. In the next section, we proceed to examine Prediction 2, which concerns the impact of providing this information to principals.

5. The Impact of Information on Employee Evaluation

The first element of Prediction 2 is that principals receiving value-added estimates should place more weight on this new information and less weight on their prior beliefs, relative to the control principals. We test this by regressing posterior evaluations (R_{it}) on prior evaluations ($R_{i,t-1}$) and teacher value-added ($V_{i,t-1}$), as shown in Equation 7. We estimate separate regressions for treatment and control groups and compare the two sets of coefficients.

$$(7)\ R_{it} = \alpha + \lambda R_{i,t-1} + \beta V_{i,t-1} + \xi_{it}$$

We find a highly significant positive effect of value-added on post-experimental evaluations for the treatment group (0.123) and a small and insignificant effect (0.017) for the control group (Table 6, Column Group 1).[26] The coefficient on prior evaluation is positive and significant for both groups. However, though the estimate on the prior evaluation for the treatment group is smaller as predicted, the difference across the two groups is not significant.

[24] For example, teachers may have taught other non-tested subjects or grade levels in prior years, or teachers may simply have had variation in their performance that is not captured by the last three years of data.

[25] Here again, we find slightly larger and more precisely estimated coefficients if we use evaluations of teachers' ability to raise student achievement instead of "overall" evaluations.

[26] The size and significance of this difference in coefficients is robust to including principal fixed effects (Column Group 2).

Including controls for teacher experience and school fixed effects has little impact on these findings (Column Group 2).

Prediction 2 of our Bayesian learning model also states that principals should place relatively more weight on value-added reports that were relatively more precise and less weight on value-added estimates for the teachers for whom they had a relatively precise prior. To test this, we interact both value-added and the principal's prior evaluation with our measures of precision, as illustrated by Equation 8.[27]

$$(8)\ R_{it} = \alpha + \lambda\left(1+\delta^{R}h_{R}\right)R_{i,t-1} + \beta\left(1+\delta^{V}h_{V}\right)V_{i,t-1} + \xi_{it}$$

As predicted, for the treatment group we find a significant positive interaction of value-added with precision and a significant negative interaction of value-added with the number of years the principal has supervised the teacher. Also, as predicted, we find a negative and marginally significant (p-value 0.11) interaction of the principal's prior evaluation with the precision of value-added and a significant positive interaction of the prior evaluation with the number of years the principal has supervised the teacher. In contrast, the interaction coefficients for the control group are much closer to zero, never even marginally significant, and sometimes of a different sign. These results are robust to including teacher experience and school fixed effects.

Thus, our findings are quite consistent with the Prediction 2 of the Bayesian learning model. Principals who receive performance data on their teachers use this information in updating their priors. They put more weight on the new data and less on their priors when the data is more precise (i.e., greater values of h_V), and less weight on new data and more on their priors when their priors are more precise (i.e., greater values of h_R).

[27] Again, our measures for precision are the inverse of the confidence interval on the value-added estimate and the number of years the teacher had been under the principal's supervision. Main effects for the precision measures are included as control variables but are omitted from Equation 8 and Table 6 to simplify exposition.

We also examine the influence of value-added on posterior evaluations separately for teachers of math, English, and both math and English. Unlike our examination of prior beliefs, where we found strong relationships for both subject areas (see Appendix Table A3), we find consistent evidence that the value-added estimates in math were more influential than those for English. We find positive significant effects of math value-added on posterior evaluations for the treatment group and not for the control group (Table 7, Column Groups 1, 3, and 5), but we find no significant effects of English value-added on posterior beliefs for either the treatment or control group (Column Groups 2, 4, and 6).

Why principals were more influenced by the value-added reports in their evaluation of math teaching is unclear. It is possible that the timing of the English exam—given in January, as opposed to math which is given in April—increased principals' concerns about the ability of the value-added methodology to measure teachers' contributions to student achievement accurately (this was discussed with principals in the training sessions). It may also be that principals were more confident in their ability to gauge instructional quality in English, and thus put less weight on the value-added estimates.

6. Information Acquisition, Worker Turnover, and Productivity

The results presented in Section 5 establish the impact of information provision on principals' subjective evaluations of work performance. However, given that the evaluations provided by principals were unofficial and carried no stakes, it is important to test whether the provision of new information actually translated into changes in personnel decisions or the quality of education provided at the school. In this section, we examine whether providing information on employee performance causes principals to gather less information via classroom observation, changes patterns of turnover, or raises student achievement.

6.1 Classroom Observation

Principals are required to evaluate all untenured and most tenured teachers formally based on one or two formal observations per year. However, they are not limited in the number of times they may observe teachers' classrooms, and our baseline survey data indicate that informal observations are quite common.[28] We first use this baseline data to test whether principals allocate time spent on teacher observation according to prior knowledge and beliefs. We then examine whether providing principals with new "hard" data on teacher performance crowded out time spent on gathering "soft" data via classroom observation during the pilot year.

To examine whether principals' allocate their time observing teachers strategically, we regress the number of observations made of teacher *i* during the year *prior* to the pilot ($O_{i,t-1}$) on the principal's baseline evaluation ($R_{i,t-1}$) the years the principal had observed the teacher at baseline ($T_{i,t-1}$), and (because between-school variation in observation frequency is quite large) set of school (principal) fixed effects (indexed *j*), as shown in Equation 9.[29]

$$(9)\ O_{i,t-1} = \beta R_{i,t-1} + \lambda T_{i,t-1} + \sum_j \pi_j D^j_{i,t-1} + \xi_{i,t-1}$$

We expect our estimate of λ to be negative, reflecting the declining value of new information as a principal learns over time. We also might expect β to be negative, since a goal of observation may be to identify ineffective teachers and provide them with constructive criticism.

These regressions provide clear support for the hypotheses that principals allocate more time to observing teachers who they know less well or who they believe are performing poorly.[30]

[28] Principals report no additional observations beyond those made for formal evaluation for just about 10 percent of teachers, and principals reported making between two and four additional "informal" observations for most teachers in the baseline year.

[29] The school fixed effects reduce our standard errors but have very little qualitative impacts on our point estimates.

[30] We drop a few teachers from these regressions for whom the principal reported formal observations but left the question on total observations blank. We *include* 15 schools (6 treatment, 9 control) in which the principal was asked about at least five teachers and reported the same number of formal and total observations in each case. These principals likely have a uniform observation policy, and all of the coefficients increase slightly if we exclude them.

An additional year of having observed the teacher in their current role reduces both formal and total observation frequency by about 0.25, while a standard deviation increase in the principal's baseline evaluation reduces formal observation frequency by 0.1 and total observation frequency by 0.2 (Table 8, Columns 1 and 3). Of course, the number of years a teacher has been observed by the principal will be positively correlated with teaching experience, and principals may simply observe experienced teachers less frequently. Controlling for overall teaching experience does cause the coefficients on years of observation to fall substantially, but they remain statistically significant (Table 8, Columns 2 and 4).

To test whether providing principals with objective performance data led them to spend less time in the classroom, we test for the significance of a treatment indicator in regressions where the dependent variable is the number of observations that the principal reported making of the teacher in the follow-up survey (O_{it}) or the difference between observations reported at follow-up (O_{it}) and baseline ($O_{i,t-1}$), as shown by Equations 10a and 10b.

$$\text{(10a)}\ O_{it} = Treat_i + \xi_t$$
$$\text{(10b)}\ O_{it} - O_{i,t-1} = Treat_i + \zeta_{it}$$

Despite the finding that principals do allocate their time strategically, we find no evidence that treatment affected principals' propensity to observe teachers in the classroom (Table 9). Estimated coefficients for the treatment indicator are positive for formal observations, negative for total observations, and statistically insignificant in all specifications. However, it should be kept in mind that, given the structure of the program and the limitations of our data, the effects would have had to take place in the short window between when reports were received in December 2007 and the end of the school year.[31]

[31] We are also limited to the sample of principals who completed the follow-up survey and the teachers who remained working during the pilot year. Repeating the specifications shown in Table 8 for this subset of principals

6.2 Employee Turnover

As discussed in Section 3, previous work suggests two main channels through which new information might impact employee turnover. First, it might have a direct impact by changing posterior beliefs, lowering evaluations of workers that would have been retained or raising evaluations of teachers that might have otherwise been let go. Second, along the lines of Baker et al. (1994), verifiable third-party information on performance may increase principals' willingness to act on available information, even if their beliefs remain unchanged.

To examine whether the new information on value-added changed patterns of turnover, we regress an indicator for whether teacher *i* is no longer employed in the same school in the year after the pilot (E_{it}) on value-added ($V_{i,t-1}$).[32] Because of possible omitted variables bias, we also estimate a specification that includes the principal's evaluation at baseline ($R_{i,t-1}$), which we know is positively correlated with value-added and is likely to be correlated with turnover.

$$\text{(11a)}\ \ E_{it} = \beta V_{i,t-1} + \xi_{it}$$
$$\text{(11b)}\ \ E_{it} = \tilde{\beta} V_{i,t-1} + \lambda R_{i,t-1} + \zeta_{it}$$

Again, we run linear regressions separately for treatment and control groups and test for differences between the groups, clustering standard errors at the school level.

We find clear evidence that providing the value-added reports did indeed cause teachers with lower value-added estimates to be more likely to exit treatment schools (Table 10). For the specification shown by Equation 11a, the coefficient on value-added is statistically significant and negative in the treatment group, and the difference between the treatment and control coefficients is significant at the 11 percent level (Column Group 1). When we include the

and teachers provides very similar coefficient estimates. Other regressions, not reported in Table 9 but available upon request, show no significant interaction between treatment status and other variables including value added, the precision of value-added, the years the principal has observed the teacher, or the principal's baseline evaluation.

[32] Results using logit specifications are quite similar to those we report here. While it would be preferable to examine involuntary and voluntary exits separately, there is no way to separate these phenomena in our data.

principal's baseline evaluation as a covariate, the value-added coefficient remains significant and negative for the treatment group and we can reject equality with the control group at the 6 percent level (Column Group 2). In addition, while the coefficients on baseline evaluation are negative for both treatment and control groups, the estimate for the control group is significantly larger in magnitude. Thus, as with our analysis of posterior evaluations, we find treatment principals putting more weight on new information and less weight on their prior beliefs.[33]

As stated above, the impact of new information on retention outcomes may be due to "direct" effects on principals' evaluations or "indirect" effects on their willingness to act, conditional on the evaluation. To distinguish these two explanations, we take advantage of the fact that only the information on *math* value-added had an impact on principals' posterior beliefs, despite the fact that math and English value-added were both strongly related to principals' priors at baseline. If the impact of information on turnover is due to direct effects on principals' posteriors, then these impacts should be driven by math value-added, not English.

This is precisely what we find. Exit is negatively related to value-added for math teachers in the treatment group but not the control group (Table 11, Column Group 1), while English value-added has very small and insignificant impacts on exit in either group (Table 11, Column Group 2).[34] This supports the interpretation that new information changed principals' posterior beliefs, and this change in beliefs in turn altered patterns of employment.

6.3 Student Achievement

[33] Including teacher experience and school fixed effects does not substantially change these results. As an additional robustness check, we also examined the probability that a teacher exited the school before the start of the pilot (i.e., between the school years 2006-2007 and 2007-2008). These "placebo tests" showed a very similar and insignificant relationship between value-added and exiting the school for treatment and control schools, while the coefficients on principals' pre-existing beliefs regarding teacher effectiveness were negative, significant, and very similar for the two groups. This supports the notion that principals make personnel decisions based on their subjective evaluations of job performance, and did this similarly in treatment and control schools prior to the start of the pilot.

[34] As robustness checks, we estimate the same specifications for teachers of both math and English (Column Groups 3 and 4), specifications that include value-added of both subjects in the same regression (Column Group 5) and specifications that control for teacher experience and school fixed effects (Column Group 6).

Providing managers with information on employee performance could be detrimental to firm productivity if that information is invalid or misleading. Thus, in addition to showing how principals incorporated information into their evaluations and personnel decisions, it is important to test whether the information led to improvements in school productivity. While the effects we find on turnover suggest higher productivity among math teachers in treatment schools, a rough calculation based on the results in Table 11 suggests increases of less than 0.01 standard deviations in students' math test scores.[35] For productivity to rise significantly, treatment principals must use this information in other ways, e.g., reallocating instructional resources (teachers' aides, mentoring, training), changing teacher assignments, or getting teachers to exert greater effort or change instructional practices. Unfortunately, we lack data to investigate these potential mechanisms.

To estimate the overall effect of the treatment on student achievement, we estimate a student-level regression of achievement gains (i.e., 2009 score minus 2008 score) on an indicator for being in a treatment school. Taking advantage of the randomization of treatment, we allow for random effects at the school and teacher level to account for the nested structure of the data and increase efficiency. We find a small but marginally significant improvement in math achievement gains of 0.024 student level standard deviations (p-value 0.16) among treatment schools (Table 12, Column 1). This estimate is insensitive to adding additional controls for teacher experience—though we lack experience data on teachers hired after the pilot—or student

[35] The coefficient on value-added in a regression of exit among math teachers in the treatment group is -.04; under the assumption that replacement teachers have zero value-added (i.e., the population average), this implies value-added in treatment schools should improve by 0.04 *teacher-level* standard deviations. However, given the variation in teacher value-added (see Table 1), this translates to 0.006 *student-level* standard deviations in test score growth. Comparing the value-added of the teachers from the pilot that actually were assigned to teach math in 2008-2009, we find differences of 0.015. This suggests that value-added may have affected grade/subject assignment as well as turnover, but still there is little evidence that teacher selection could fully explain the change in test scores.

characteristics and grade level covariates (Column 2 and 3).[36] As expected, we find no evidence of significant improvements in English and the treatment coefficient ranges from -0.01 to 0.01 depending on our control variables (Columns 7 to 9).

Due to turnover or reassignment to non-tested grades or subjects, fewer than half of the students included in these regressions were taught by a teacher who was working in these schools during the pilot study. When we limit our sample to students taught by teachers in the pilot, the point estimate on treatment rises to 0.04 in math and has a p-value of 0.10 (Column 4), while the estimate in English remains quite close to zero (Column 10). Thus, math achievement gains produced by teachers in the pilot were significantly higher in the treatment schools.

The effect on math achievement growth is considerably higher than we would have predicted if the only channel for increased productivity was selection of who remained teaching in treatment and control school classrooms. Including the principal's evaluation at baseline increases the treatment effect estimate to 0.057 (Column 5), and adding the teacher's value-added shrinks the treatment coefficient to 0.044. This is consistent with our results on turnover, i.e., new information caused the retention (dismissal) of some teachers who had high (low) value-added but of whom the principal held a low (high) baseline opinion, but it is evident that selection on these attributes cannot fully explain the impact of treatment on math test score gains.

Overall, we would stress that the positive impact of the NYC pilot on productivity is very small compared with increased school quality (Hastings and Weinstein 2008), reduced class size (Krueger, 1999), or severe accountability pressure (Rouse et al., 2007), which have been found

[36] Student characteristics include: prior test score, prior test score interacted with grade level, prior test score in the other subject (e.g., reading when predicting math gains), student gender, racial/ethnic subgroup, English language learner status, special education status, and eligibility for free or reduced price lunch. Grade level covariates include grade fixed effects interacted with the grade configuration of the school (e.g., grade 6 students in middle schools).

to raise student achievement by 0.15 to 0.25 standard deviations. However, the provision of value-added estimates on pre-existing student examinations is a fairly low cost intervention.

7. Conclusion

We study a pilot experiment providing "value-added" estimates of teacher performance to New York City school principals to learn about how managers evaluate employee job performance and how objective performance data influences this process. We present a number of empirical facts consistent with a simple Bayesian model where principals learn over time with imperfect information. First, the positive relationship between value-added and principals' prior beliefs about teacher effectiveness is stronger when value-added measures are more precisely estimated or when the principal has supervised the teacher for a longer period of time—our proxy for the precision of principals' priors. Second, principals change their evaluations of teachers in response to the value-added estimates, and the impact of this new information rises with its precision and falls with the precision of the principals' prior beliefs. Thus, value-added estimates are capturing a dimension of teacher performance valued by school principals and are providing new information to principals on this performance dimension.

Importantly, the provision of new information also had real impacts on personnel decisions and productivity. Teachers with lower value-added estimates were more likely to exit their schools, and students show small, marginally significant improvements in test scores. Importantly, comparisons of impacts across subject areas (math vs. English) suggest that these changes arose via the influence of new information on principals' posterior beliefs. Despite the fact that objective performance data in both English and math bear strong relationships with principals' priors, the changes in principals' posterior beliefs, patterns of teacher turnover, and student test score improvements all are driven by the provision of performance data in math.

These results support the notion that standardized teacher performance data are useful to principals in managing their schools, and adds to the growing literature on the benefits of governmental provision of information where acquisition may be costly.[37] While our results provide modest support for moving school districts towards the use of student achievement in evaluating teachers, there are important caveats in drawing policy conclusions from this pilot program. First, studies of accountability systems provide ample evidence of gaming and cheating on high stakes performance measures, particularly standardized exams whose format and content are well-known. In a recent paper, Neal (2010) highlights this issue and offers concrete suggestion for creating measures of teacher performance that avoid these pitfalls. Second, the notion that value-added estimates cannot capture all of the dimensions of performance valued by principals and society in general is supported by many pieces of evidence, including our baseline survey and the modest size of the correlation between value-added and principals' prior beliefs. While current systems of teacher evaluation have been criticized for ignoring teachers' impacts on student achievement, a system that focus heavily on value-added measures may risk placing too little weight on other aspects of teaching that increase student and social welfare.

Motivated in part by research in economics on the importance of teachers in educational production, the federal government has encouraged states and school districts to use student achievement growth to measure teacher effectiveness as part of the incentives built into its $4.3 billion Race to the Top Fund. Encouraged by federal incentives, many states are moving quickly

[37] Recent examples include restaurant hygiene (Jin and Leslie, 2003), hospital quality (Dranove et al., 2003; Hibbard et al., 2005; Chassin et al., 2010), and school quality (Hastings and Weinstein, 2008; Andrabi et al., 2009). Of course, the fact that principals view the information as useful and incorporate into their evaluations does not guarantee that value-added captures a dimension of productivity valued by parents or society. While we cannot directly address the validity of the value-added measures provided to principals in the pilot program, Chetty et al. (2011) present evidence that value-added measures similar to those used in the pilot do appear to be free of bias and capture a dimension of productivity that is related to students' future educational and labor market outcomes.

to establishing new regulations for how teachers are evaluated and compensated, and researchers will have great opportunities to study the impact of these new measures and new programs on productivity in public education.

References

Aaronson, D., Barrow, L. & Sander, W. (2007) “Teachers and Student Achievement in the Chicago Public High Schools,” Journal of Labor Economics, 25(1): 95-135.

Altonji, J.G. and Pierret, C.R. (2001) “Employer Learning and Statistical Discrimination,” Quarterly Journal of Economics, 116(1): 313-350.

Andrabi, T., Das, J., and Khwaja, A.I. (2009), “Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets,” Unpublished Working Paper.

Armor, David, Patricia Conry-Oseguera, Millicent Cox, Nicelma King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Santa Monica, CA: Rand Corporation.

Baker, G.P., Jensen, M.C., and Murphy, K.J. (1988) “Compensation and Incentives: Practice vs. Theory,” Journal of Finance, 43(3): 593-616.

Baker, G.P., Gibbons, R. and Murphy, K.J. (1994) “Subjective Performance Measures in Optimal Incentive Contracts,” Quarterly Journal of Economics, 109 (4): 1125-1156.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd , H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., and Shepard, L.A. (2010) “Problems with the Use of Student Test Scores to Evaluate Teachers,” Economic Policy Institute Paper #278.

Bloom, N., Eifert, B., Mahajan, A. McKenzie, D., and Roberts, J. (2010) “Does Management Matter? Evidence from India,” Unpublished Manuscript.

Boyd, D., Grossman, P., Lankford, H., Loeb, S. & Wyckoff, J. (2007) “Who Leaves? Teacher Attrition and Student Achievement,” Unpublished Working Paper.

Chassin, M.R., Loeb, J.M., Schmaltz, S.P., Wachter, R.M. (2010), “Accountability Measures – Using Measurement to Promote Quality Improvement,” New England Journal of Medicine, 363(7): 683-688.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Whitmore Schanzenbach, D., Yagan, D. (2010) “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR,” NBER Working Paper No. 16381

Chetty, R., Friedman, J.N., Rockoff, J.E. (2011) “The Impact of Teacher Value Added on Student Outcomes in Adulthood,” Unpublished Manuscript.

Clark, D., Martorell, P., and Rockoff, J.E. (2009) “School Principals and School Performance,” Calder Center Working Paper #38.

Corcoran, S. (2010) “Can Teachers be Evaluated by their Students’ Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice,” Working Paper, Annenberg Institute for School Reform.

DeGroot, M. (1970) Optimal Statistical Decisions. New York: McGraw Hill.

Dranove, D., Kessler, D., McClellan, M. and Satterthwaite, M. (2003) “Is More Information Better? The Effects of ‘Report Cards’ on Health Care Providers,” Journal of Political Economy, 111(3): 555-588.

Farber, H.S. and Gibbons, R. (1996) “Learning and Wage Dynamics,” Quarterly Journal of Economics, 111(4): pp. 1007-1047.

Gibbons R. and Katz, L. (1991) “Layoffs and Lemons,” Journal of Labor Economics 9(4): 351-380.

Gibbons, R., Katz, L., Lemieux, T., and Parent, D. (2005) “Comparative Advantage, Learning, and Sectoral Wage Determination,” Journal of Labor Economics, 23(4): 681-724.

Goldhaber, D. and Hansen, M. (2009) “Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions,” University of Washington Center on Reinventing Public Education Working Paper 2009_2.

Gordon, R., Kane, T., & Staiger, D. (2006) The Hamilton Project: Identifying Effective Teachers Using Performance on the Job. Washington, DC: The Brookings Institution.

Grissom, J. and Loeb, S. (2009) “Triangulating Principal Effectiveness: How Perspectives of Parents, Teachers and Assistant Principals Identify the Central Importance of Managerial Skills,” Calder Working Paper 35.

Harris, D.N. and Sass, T.R. (2006) “Value-added Models and the Measurement of Teacher Quality,” University of Florida Working Paper.

Harris, D.N. and Sass, T.R. (2009) “What Makes for a Good Teacher and Who Can Tell?” Calder Center Working Paper #30.

Harris, M. and Holmstrom, B. (1982) “A Theory of Wage Dynamics,” The Review of Economic Studies, 49(3): 315-333.

Hastings, J.S. and Weinstein J.M. (2008) “Information, School Choice, and Academic Achievement: Evidence from Two Experiments,” Quarterly Journal of Economics, 123(4): 1373-1414.

Hanushek, E.A. and Rivkin S.G. (2010) “Generalizations about Using Value-Added Measures of Teacher Quality,” American Economic Review, Papers and Proceedings 100(2): 267-271.

Hibbard, J.H., Stockard, J., and Tusler, M. (2005) “Hospital Performance Reports: Impact on Quality, Market Share, and Reputation,” Health Affairs, 24(4): 1150-1160.

Holmstrom, B. and Milgrom, P. (1991) “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design,” Journal of Law, Economics, and Organization, 7(Sp): 24-52.

Jacob, B.A. (2007) “The Demand Side of the Teacher Labor Market,” Unpublished Manuscript, University of Michigan.

Jacob, B.A. (2010) “Do Principals Fire the Worst Teachers?” NBER Working Paper 15715.

Jacob, B.A., and Lefgren, L.J. (2008) “Principals as Agents: Subjective Performance Measurement in Education” Journal of Labor Economics 26(1): 101-136.

Jin, G.Z. and Leslie, P. (2003) “The Effect Of Information On Product Quality: Evidence From Restaurant Hygiene Grade Cards,” Quarterly Journal of Economics, 118(2): 409-451.

Jovanovic, B. (1979) “Job Matching and the Theory of Turnover,” Journal of Political Economy, 87(5): 972-990.

Kane, T.J. and Staiger, D.O. (2008) “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation” NBER Working Paper #14607.

Koedel, C. and Betts, J.R. (2009) “Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique,” Education Finance and Policy 6(1): 18-42.

Krueger, A. (1999) “Experimental Estimates of Education Production Functions,” Quarterly Journal of Economics, 114(2): 497-532.

Macleod, W.B. (2003) “Optimal Contracting with Subjective Evaluation,” American Economic Review, 93(1): 216-240.

Medoff, J.L. and Abraham, K.G. (1980) “Experience, Performance, and Earnings,” Quarterly Journal of Economics, 95(4): 703-736.

Murnane, Richard J. (1975) *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Balinger.

Murphy, K.J. (1986) “Incentives, Learning, and Compensation: A Theoretical and Empirical Investigation of Managerial Labor Contracts,” RAND Journal of Economics 17(1): pp. 59-76.

Murphy, K.J. (1992) “Performance Measurement and Appraisal: Motivating Managers to Identify and Reward Performance,” in W.J.J. Burns (Ed.), Performance Measurement, Evaluation, and Incentives, Boston, MA: Harvard Business School Press pp. 37-62.

Murphy, K.J. and Oyer, P. (2001) “Discretion in Executive Incentive Contracts: Theory and Evidence,” Unpublished Manuscript.

Neal, Derek (2011) “The Design of Performance Pay in Education,” NBER Working Paper 16710.

New York Times (2008), “New York Measuring Teachers by Test Scores,” by Jennifer Medina, January 21, 2008, page A1.

Oyer, P. and Schaeffer, S. (2011) “Personnel Economics: Hiring and Incentives” in Orley Ashenfelter and David Card (Eds.), Handbook of Labor Economics, Great Britain, North Holland, pp. 1769-1823.

Prendergast, C. and Topel, R. (1993) “Discretion and Bias in Performance Evaluation,” European Economic Review, 37(1): 355-365.

Rivkin, S.G., Hanushek, E. A. & Kain, J. (2005) “Teachers, Schools, and Academic Achievement,” Econometrica, 73(2): 417–458.

Rockoff, J. E. (2004) “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data,” American Economic Review, 94(2): 247-252.

Rockoff, J.E. and Turner, L.J. (2010) “Short Run Impacts of Accountability on School Quality,” American Economic Journal: Economic Policy, 2(4): 119-147.

Rothstein, J. (2009) “Student Sorting and Bias in Value-added Estimation: Selection on Observables and Unobservables,” Education Finance and Policy, 4(4): 537-571.

Spence, M. (1973) “Job Market Signaling,” Quarterly Journal of Economics, 87(3): 355-374.

Staiger, D.O. and Rockoff, J.E. (2010) “Searching for Effective Teachers with Imperfect Information,” Journal of Economic Perspectives, Summer 2010.

Todd, P.E. and Wolpin, K.I. (2007) “On the Specification and Estimation of the Production Function for Cognitive Achievement,” Economic Journal, 113(1): 3-33.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009) The Widget Effect. Brooklyn, NY: The New Teacher Project.

Figure 1: Timeline of Teacher Data Initiative Pilot Program

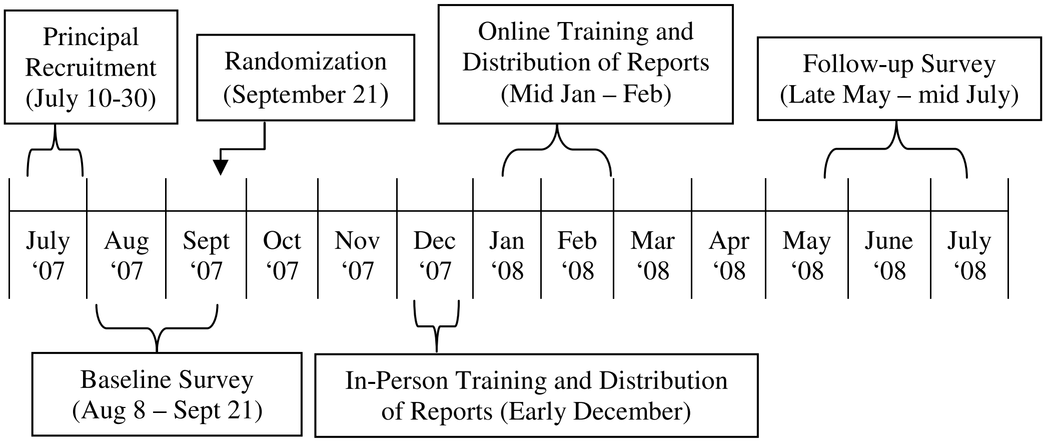
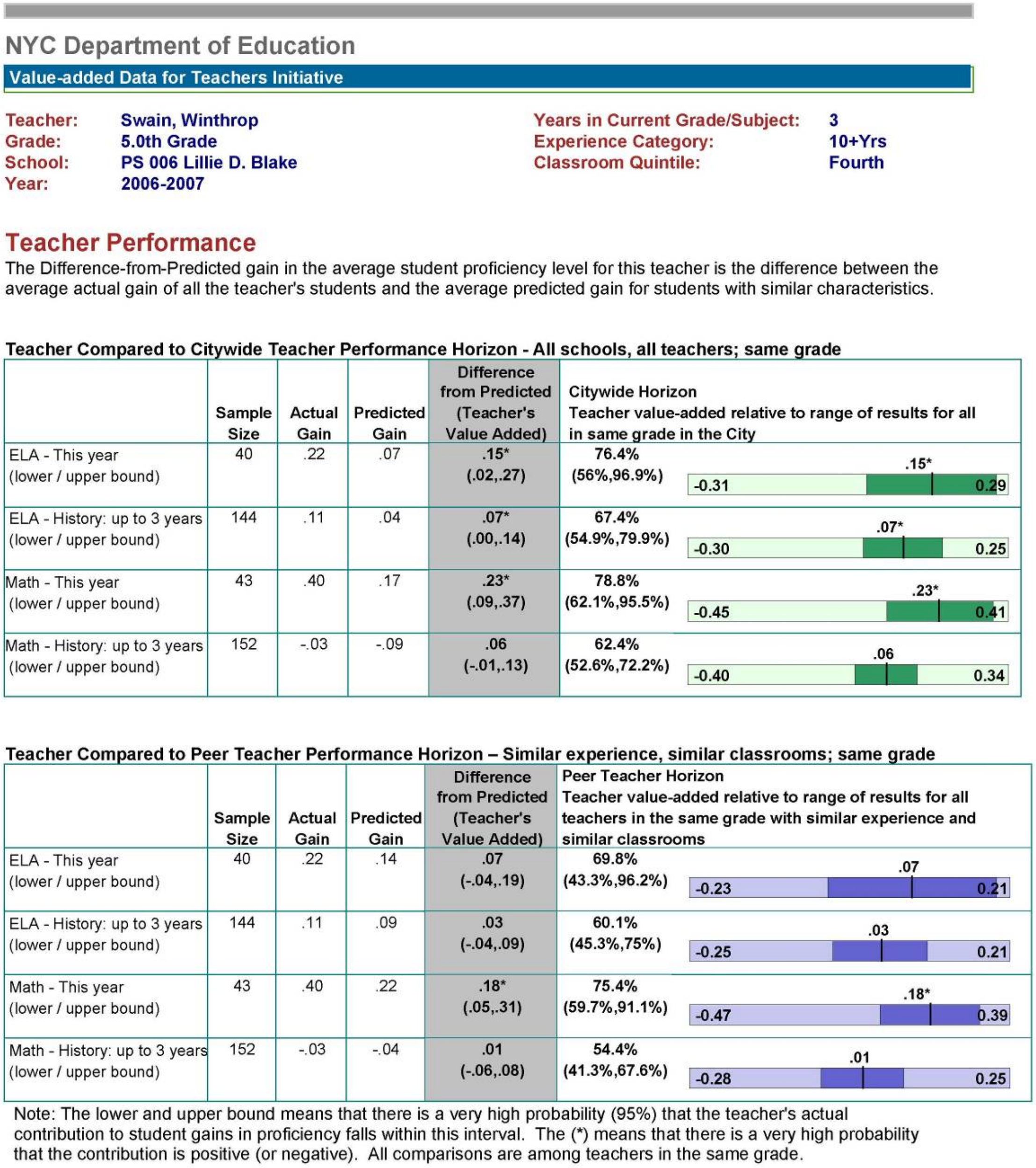

Principal Recruitment (July 10-30)
Randomization (September 21)
Online Training and Distribution of Reports (Mid Jan – Feb)
Follow-up Survey (Late May – mid July)
July '07
Aug '07
Sept '07
Oct '07
Nov '07
Dec '07
Jan '08
Feb '08
Mar '08
Apr '08
May '08
June '08
July '08
Baseline Survey (Aug 8 – Sept 21)
In-Person Training and Distribution of Reports (Early December)

Figure 2: Sample Value Added Report

Page 1



NYC Department of Education

Value-added Data for Teachers Initiative

Teacher:	Swain, Winthrop	Years in Current Grade/Subject:	3
Grade:	5.0th Grade	Experience Category:	10+Yrs
School:	PS 006 Lillie D. Blake	Classroom Quintile:	Fourth
Year:	2006-2007		

Teacher Performance

The Difference-from-Predicted gain in the average student proficiency level for this teacher is the difference between the average actual gain of all the teacher's students and the average predicted gain for students with similar characteristics.

Teacher Compared to Citywide Teacher Performance Horizon - All schools, all teachers; same grade

	Sample Size	Actual Gain	Predicted Gain	Difference from Predicted (Teacher's Value Added)	Citywide Horizon Teacher value-added relative to range of results for all in same grade in the City	
ELA - This year (lower / upper bound)	40	.22	.07	.15* (.02,.27)	76.4% (56%,96.9%)	.15* -0.31 0.29
ELA - History: up to 3 years (lower / upper bound)	144	.11	.04	.07* (.00,.14)	67.4% (54.9%,79.9%)	.07* -0.30 0.25
Math - This year (lower / upper bound)	43	.40	.17	.23* (.09,.37)	78.8% (62.1%,95.5%)	.23* -0.45 0.41
Math - History: up to 3 years (lower / upper bound)	152	-.03	-.09	.06 (-.01,.13)	62.4% (52.6%,72.2%)	.06 -0.40 0.34

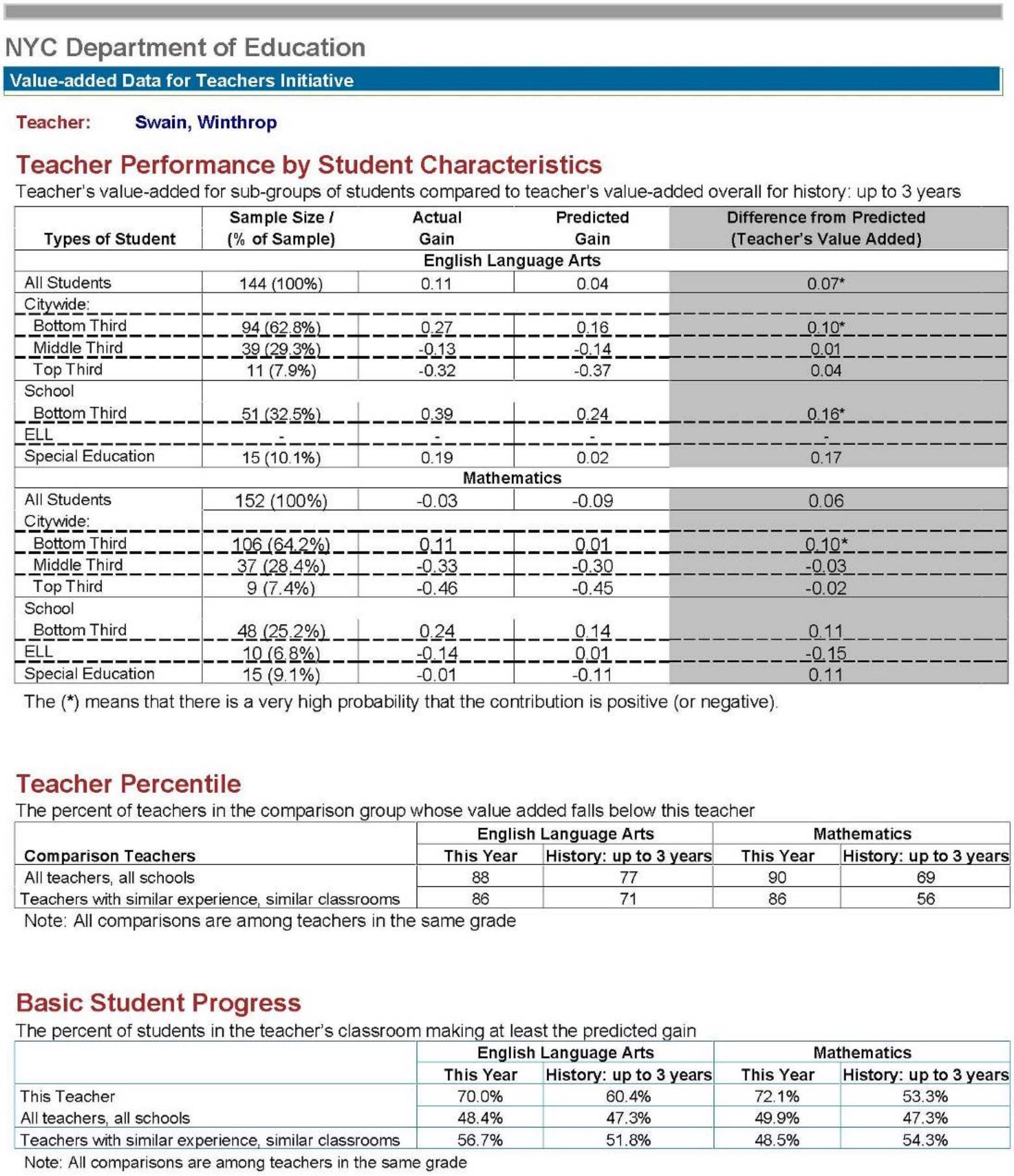
Teacher Compared to Peer Teacher Performance Horizon – Similar experience, similar classrooms; same grade

	Sample Size	Actual Gain	Predicted Gain	Difference from Predicted (Teacher's Value Added)	Peer Teacher Horizon Teacher value-added relative to range of results for all teachers in the same grade with similar experience and similar classrooms	
ELA - This year (lower / upper bound)	40	.22	.14	.07 (-.04,.19)	69.8% (43.3%,96.2%)	.07 -0.23 0.21
ELA - History: up to 3 years (lower / upper bound)	144	.11	.09	.03 (-.04,.09)	60.1% (45.3%,75%)	.03 -0.25 0.21
Math - This year (lower / upper bound)	43	.40	.22	.18* (.05,.31)	75.4% (59.7%,91.1%)	.18* -0.47 0.39
Math - History: up to 3 years (lower / upper bound)	152	-.03	-.04	.01 (-.06,.08)	54.4% (41.3%,67.6%)	.01 -0.28 0.25

Note: The lower and upper bound means that there is a very high probability (95%) that the teacher's actual contribution to student gains in proficiency falls within this interval. The (*) means that there is a very high probability that the contribution is positive (or negative). All comparisons are among teachers in the same grade.

Figure 2: Sample Value Added Report

Page 2



NYC Department of Education

Value-added Data for Teachers Initiative

Teacher: **Swain, Winthrop**

Teacher Performance by Student Characteristics

Teacher's value-added for sub-groups of students compared to teacher's value-added overall for history: up to 3 years

Types of Student	Sample Size / (% of Sample)	Actual Gain	Predicted Gain	Difference from Predicted (Teacher's Value Added)
		English Language Arts		
All Students	144 (100%)	0.11	0.04	0.07*
Citywide:				
Bottom Third	94 (62.8%)	0.27	0.16	0.10*
Middle Third	39 (29.3%)	-0.13	-0.14	0.01
Top Third	11 (7.9%)	-0.32	-0.37	0.04
School				
Bottom Third	51 (32.5%)	0.39	0.24	0.16*
ELL	-	-	-	-
Special Education	15 (10.1%)	0.19	0.02	0.17
		Mathematics		
All Students	152 (100%)	-0.03	-0.09	0.06
Citywide:				
Bottom Third	106 (64.2%)	0.11	0.01	0.10*
Middle Third	37 (28.4%)	-0.33	-0.30	-0.03
Top Third	9 (7.4%)	-0.46	-0.45	-0.02
School				
Bottom Third	48 (25.2%)	0.24	0.14	0.11
ELL	10 (6.8%)	-0.14	0.01	-0.15
Special Education	15 (9.1%)	-0.01	-0.11	0.11

The (*) means that there is a very high probability that the contribution is positive (or negative).

Teacher Percentile

The percent of teachers in the comparison group whose value added falls below this teacher

	English Language Arts		Mathematics	
Comparison Teachers	This Year	History: up to 3 years	This Year	History: up to 3 years
All teachers, all schools	88	77	90	69
Teachers with similar experience, similar classrooms	86	71	86	56

Note: All comparisons are among teachers in the same grade

Basic Student Progress

The percent of students in the teacher's classroom making at least the predicted gain

	English Language Arts		Mathematics	
	This Year	History: up to 3 years	This Year	History: up to 3 years
This Teacher	70.0%	60.4%	72.1%	53.3%
All teachers, all schools	48.4%	47.3%	49.9%	47.3%
Teachers with similar experience, similar classrooms	56.7%	51.8%	48.5%	54.3%

Note: All comparisons are among teachers in the same grade

Table 1: Summary Statistics on Teacher Level Variables at Baseline

	Control	Treatment	P-value on Difference
Number of Teachers	1,214	1,337	
Principal's Rating (Scale from 1 to 6)			
Overall	4.32	4.31	0.90
	(1.12)	(1.12)	
Math Instruction	4.21	4.23	0.85
	(1.09)	(1.13)	
ELA Instruction	4.20	4.19	0.89
	(1.04)	(1.13)	
Observations Made by Principal Last Year			
Formal	4.19	4.19	0.87
	(1.08)	(1.12)	
Total	2.21	2.23	0.61
	(1.27)	(1.25)	
Value-added Estimates			
Math, Multi-year, Citywide	0.00	0.01	0.94
	(0.15)	(0.16)	
ELA, Multi-year, Citywide	-0.01	0.00	0.23
	(0.10)	(0.11)	
Precision of Value-added Estimates			
Math, Multi-year, Citywide	1.92	1.94	0.87
	(0.99)	(1.01)	
ELA, Multi-year, Citywide	1.78	1.77	0.83
	(1.02)	(0.99)	
Total Teaching Experience			
None (First Year was 2006-2007)	0.09	0.11	0.27
One Year	0.12	0.11	0.66
Two Years	0.12	0.11	0.95
Three Years	0.08	0.11	0.07
Four Years	0.08	0.07	0.44
Five to Nine Years	0.27	0.26	0.64
Ten or More Years	0.24	0.23	0.54
Years Principal Observes Teacher in Current Role	1.99	1.96	0.71
	(0.88)	(0.87)	

Note: P-values indicate the statistical significance of the difference between treatment and control schools. Standard deviations in parentheses. Teachers for whom the principal reported more than 10 total observations made in the last year are given a value of 11. Precision of Value-added is the inverse of the 95 percent confidence interval for the value-added estimate. Years Principal Observes Teacher is calculated as the minimum of principal experience in the current school and the number of years of data used in the teacher's value-added estimate.

Table 2: Comparisons of Baseline Characteristics for Treatment vs. Control Groups and Study Sample vs. Population

	Treatment vs. Control (Baseline)			*Treatment vs. Control (Follow-up)*			*Population vs. Sample*		
	Control Mean	Treatment Mean	P-value on Difference	Control Mean	Treatment Mean	P-value on Difference	Population Mean	Sample Mean	P-value on Difference
Number of Principals/Schools	111	112		94	84		1092	223	
Total Enrollment	705	718	0.79	715	725	0.85	660	712	0.03
Principal Characteristics (Spring 2007)									
Years of Experience as Principal (in School)	3.3	3.2	0.81	3.2	3.2	0.98	3.6	3.3	0.27
Years of Experience as Assistant Principal	2.4	2.7	0.40	2.4	2.8	0.43	2.7	2.5	0.41
Years of Experience as Teacher	6.8	7.6	0.36	6.6	7.8	0.18	6.2	7.2	0.01
Years of Experience in School (Any Position)	4.4	5.1	0.31	4.4	4.8	0.53	5.2	4.7	0.17
Principal Age	48.0	48.8	0.50	48.2	49.9	0.16	48.8	48.4	0.41
Principal is Black or Hispanic	48.6%	41.9%	0.32	47.9%	44.1%	0.61	47.7%	45.3%	0.47
Principal is Female	81.1%	77.7%	0.53	80.9%	77.4%	0.57	73.4%	79.4%	0.05
Student Characteristics (Grades 4-8, 2007)									
Math Test Pass Rate	68.5%	69.7%	0.66	70.2%	70.2%	1.00	67.4%	69.1%	0.23
English Test Pass Rate	50.4%	52.8%	0.38	51.1%	53.1%	0.50	52.0%	51.6%	0.80
On Free Lunch	87%	85.4%	0.64	85.7%	85.0%	0.81	84.9%	86.1%	0.35
English Language Learners	15%	13.3%	0.30	14.7%	13.0%	0.30	11.5%	14.0%	0.00
In Special Education	10%	9.8%	0.79	9.2%	10.2%	0.40	13.8%	9.6%	0.00
Black or Hispanic	72%	72.8%	0.91	71.9%	72.9%	0.83	75.6%	72.5%	0.11
Teacher Survey Results (Spring 2007)									
The Principal…									
Visits Classrooms to Observe the Quality of Teaching	0.071	-0.03	0.44	0.091	-0.026	0.44	0.00	0.03	0.65
Gives Me Regular and Helpful Feedback	-0.047	-0.069	0.86	-0.026	-0.101	0.59	0.00	-0.05	0.47
Places a High Priority on the Quality of Teaching	-0.027	0.004	0.80	-0.031	0.034	0.64	0.00	-0.02	0.72
Teachers in this School…									
Use Student Data to Improve Instructional Decisions	0.096	0.077	0.87	0.093	0.058	0.80	0.00	0.00	0.99
Receive Training in the Use of Student Data	0.036	0.022	0.90	0.049	0.012	0.79	0.00	-0.01	0.90

Note: P-values indicate the statistical significance of the difference between the two groups of principals/schools. Teacher survey variables have been normalized using schools city-wide to have mean zero and standard deviation one. One control and three treatment schools are missing survey data.

Table 3: Principals' Pre-experimental Performance Evaluations and Value-Added

	(1)	(2)	(3)	(4)	(5)	(6)
Value-added, Multi-year, Peer	0.225**			0.206**	0.153**	0.257**
	(0.023)			(0.051)	(0.038)	(0.058)
Value-added, Single-year, Peer		0.207**		0.022		
		(0.022)		(0.050)		
Value-added, Multi-year, Citywide			0.213**		0.099**	-0.036
			(0.023)		(0.037)	(0.070)
Teacher Experience Fixed Effects						√
R-squared	0.05	0.04	0.05	0.05	0.06	0.08
Sample Size	2,551	2,551	2,551	2,551	2,551	2,551

Note: The dependent variable is the principal's overall evaluation of the teacher, standardized to have mean zero and standard deviaiton one. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 4: Performance Evaluations and Information Precision

	(1)	(2)	(3)
Value-added	0.079**	0.102**	0.102**
	(0.027)	(0.028)	(0.028)
Value-added * Estimate Precision	0.142**	0.151**	0.150**
	(0.027)	(0.024)	(0.024)
Teacher Experience and School FE		√	√
Years of Value-added Data FE			√
R-squared	0.09	0.35	0.35
Sample Size	2,551	2,551	2,551

Note: Value-added refers to estimates based on up to three years of data and comparisons with peers. Precision is measured as the inverse of the standard-error of the value-added estimate, normalized to have a minimum value of zero and a standard deviation of one. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 5: Value-Added and the Precision of Performance Evaluations

	(1)	(2)	(3)
Principal's Overall Evaluation	0.151**	0.103*	0.056
	(0.032)	(0.046)	(0.084)
Principal's Overall Evaluation	0.011+	0.016+	0.034*
* Years of Experience as Principal	(0.007)	(0.008)	(0.016)
Limited to Teachers with <10 Years Experience		√	
Limited to Teachers with <5 Years Experience			√
R-squared	0.08	0.06	0.08
Sample Size	1,215	815	363

Note: The dependent variable is the teacher's value-added estimate (combining math and English) based on three years of data and comparisons to peers. Only teachers with three years of data used in their value-added estimate are included in the sample. All specifications include a control for years of experience as principal, though in no regression is this coefficient statistically significant. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 6: The Impact of Value-added Information on Performance Evaluations

	(1)			(2)			(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	0.125**	0.017	*0.108*	0.151**	0.030	*0.121*	0.089*	0.034	*0.055*	0.142**	0.039	*0.103*
	(0.032)	(0.023)	*[p=0.006]*	(0.033)	(0.024)	*[p=0.003]*	(0.042)	(0.028)	*[p=0.276]*	(0.048)	(0.035)	*[p=0.083]*
Overall Evaluation, Pre-experiment	0.792**	0.821**	*-0.029*	0.720**	0.768**	*-0.048*	0.761**	0.759**	*0.002*	0.661**	0.735**	*-0.074*
	(0.033)	(0.034)	*[p=0.541]*	(0.042)	(0.038)	*[p=0.397]*	(0.071)	(0.065)	*[p=0.983]*	(0.075)	(0.076)	*[p=0.488]*
Estimate Precision												
* Value-added							0.076*	-0.012	*0.088*	0.052	-0.008	*0.06*
							(0.031)	(0.035)	*[p=0.06]*	(0.035)	(0.031)	*[p=0.2]*
* Overall Evaluation							-0.051+	0.005	*-0.056*	-0.046	-0.002	*-0.044*
							(0.029)	(0.034)	*[p=0.21]*	(0.030)	(0.037)	*[p=0.356]*
Years Principal Observes Teacher												
* Value-added							-0.089*	-0.013	*-0.076*	-0.085*	-0.006	*-0.079*
							(0.039)	(0.034)	*[p=0.142]*	(0.040)	(0.038)	*[p=0.152]*
* Overall Evaluation							0.121**	0.051	*0.070*	0.140**	0.030	*0.110*
							(0.033)	(0.037)	*[p=0.158]*	(0.037)	(0.043)	*[p=0.053]*
Joint Test of Equal Coefficients			*7.35*			*9.88*			*11.86*			*17.89*
			[p=0.03]			*[p=0.01]*			*[p=0.16]*			*[p=0.02]*
Experience Controls				√	√					√	√	
School Fixed Effects				√	√					√	√	
R-squared	0.56	0.58		0.69	0.68		0.57	0.58		0.70	0.68	
Sample Size	759	814		759	814		759	814		759	814	

Note: The dependent variable is the principal's overall evaluation of the teacher in the follow-up survey. Value-added refers to estimates based on up to three years of data and comparisons with peers. Precision is measured as the inverse of the confidence interval of the value-added estimate, normalized to have a minimum value of zero and a standard deviation of one. Years principal has supervised the teacher is equal to the minimum of the years of data used to construct the valued added estimate and the principals years of experience in the school. All specifications control for teacher experience fixed effects. Standard errors (in parentheses) are clustered by school; p-values on the test of differences in brackets. **p < 0.01, *p<0.05, +p<0.1.

Table 7: Impact of Value-added on Performance Evaluation, by Subject

	Math			English		
	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added, Math	0.147**	0.003	*0.144*			
	(0.032)	(0.028)	*[p=0.001]*			
Value-added, English				0.031	0.029	*0.002*
				(0.034)	(0.027)	*[p=0.963]*
Overall Evaluation, Pre-experiment	0.772**	0.819**	*-0.047*	0.818**	0.813**	*0.005*
	0.037	0.035	*[p=0.356]*	0.04	0.041	*[p=0.93]*
R-squared	0.57	0.57		0.55	0.55	
Sample Size	616	631		580	607	
	Math Only			English Only		
	(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added, Math	0.192**	0.036	*0.156*			
	(0.054)	(0.051)	*[p=0.036]*			
Value-added, English				-0.088	-0.003	*-0.085*
				(0.075)	(0.08)	*[p=0.439]*
Overall Evaluation, Pre-experiment	0.765**	0.845**	*-0.08*	0.915**	0.842**	*0.073*
	(0.068)	(0.053)	*[p=0.354]*	(0.045)	(0.086)	*[p=0.453]*
R-squared	0.60	0.66		0.63	0.57	
Sample Size	156	161		108	129	
	Both Math & English					
	(5)			(6)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added, Math	0.121**	-0.028	*0.149*			
	(0.035)	(0.027)	*[p=0.001]*			
Value-added, English				0.025	0.034	*-0.009*
				(0.036)	(0.031)	*[p=0.85]*
Overall Evaluation, Pre-experiment	0.773**	0.804**	*-0.031*	0.798**	0.795**	*0.003*
	(0.043)	(0.046)	*[p=0.623]*	(0.047)	(0.046)	*[p=0.964]*
R-squared	0.55	0.53		0.54	0.53	
Sample Size	452	458		452	458	

Note: The dependent variable is the principal's evaluation of a teacher's overall effectiveness in the follow-up survey. Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 8: Teacher Observation by Principals at Baseline

	Formal Observations		Total Observations	
	(1)	(2)	(4)	(5)
Years Principal Observes Teacher	-0.272**	-0.090**	-0.221**	-0.076+
	(0.027)	(0.027)	(0.038)	(0.039)
Overall Performance Evaluation	-0.116**	-0.107**	-0.228**	-0.222**
	(0.022)	(0.022)	(0.040)	(0.041)
Teacher Experience (No Experience Omitted)				
1 Years Experience		-0.045		-0.268+
		(0.082)		(0.137)
2 Years Experience		-0.150+		-0.314*
		(0.088)		(0.128)
3 Years Experience		-0.353**		-0.413**
		(0.113)		(0.154)
4 Years Experience		-0.649**		-0.642**
		(0.108)		(0.148)
5-9 Years Experience		-0.806**		-0.670**
		(0.100)		(0.139)
10+ Years Experience		-0.884**		-0.815**
		(0.102)		(0.148)
R-squared	0.71	0.77	0.89	0.89
Sample Size	2,487	2,487	2,487	2,487

Note: All regressions include school fixed effects. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 9: Impact of Value-Added Information on Classroom Observation

	Formal Observations		Total Observations	
	Levels	Changes	Levels	Changes
	(1)	(2)	(3)	(4)
Treatment School	0.064	0.165	-0.175	-0.210
	(0.161)	(0.126)	(0.513)	(0.525)
R-squared	0.00	0.01	0.00	0.00
Sample Size	1,523	1,523	1,520	1,520

Note: Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 10: Impact of Value-added Information on a Teachers' Propensity to Exit the School

	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	-0.026*	-0.001	*-0.025*	-0.021+	0.009	*-0.03*
	(0.012)	(0.010)	*[p=0.11]*	(0.012)	(0.011)	*[p=0.065]*
Overall Evaluation, Pre-experiment				-0.018	-0.051**	*0.033*
				(0.013)	(0.015)	*[p=0.097]*
Teacher Experience and School FE						
R-squared	0.01	0.00		0.01	0.02	
Sample Size	1,103	1,032		1,103	1,032	

Note: Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school; p-values on the test of differences in brackets. **p < 0.01, *p<0.05, +p<0.1.

Table 11: Impact of Value-added on Propensity to Exit the School, by Subject

	Teachers of Math			Teachers of English		
	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added in Math	-0.031*	0.006	*-0.037*			
	(0.013)	(0.012)	*[p=0.037]*			
Value-added in English				0.000	0.003	*-0.003*
				(0.014)	(0.013)	*[p=0.875]*
Overall Evaluation, Pre-experiment	-0.020	-0.041**	*0.021*	-0.013	-0.052**	*0.039*
	(0.014)	(0.015)	*[p=0.306]*	(0.015)	(0.016)	*[p=0.076]*
Sample Size	936	844		872	819	
	Teachers of Math and English					
	(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added in Math	-0.036*	0.003	*-0.039*			
	(0.016)	(0.015)	*[p=0.076]*			
Value-added in English				-0.000	-0.011	*0.011*
				(0.028)	(0.015)	*[p=0.729]*
Overall Evaluation, Pre-experiment	-0.008	-0.037*	*0.029*	-0.017	-0.035*	*0.018*
	(0.015)	(0.017)	*[p=0.201]*	(0.016)	(0.017)	*[p=0.441]*
Sample Size	705	631		705	631	
	Teachers of Math and English					
	(5)			(6)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added in Math	-0.043*	0.010	*-0.053*	-0.042*	-0.006	*-0.036*
	(0.017)	(0.014)	*[p=0.016]*	(0.018)	(0.02)	*[p=0.181]*
Value-added in English	0.016	-0.015	*0.031*	0.015	-0.001	*0.016*
	(0.028)	(0.015)	*[p=0.329]*	(0.028)	(0.019)	*[p=0.636]*
Overall Evaluation, Pre-experiment	-0.011	-0.036*	*0.025*	-0.007	-0.031	*0.024*
	(0.015)	(0.017)	*[p=0.27]*	(0.016)	(0.02)	*[p=0.349]*
Teacher Experience and School FE				√	√	
Sample Size	705	631		705	631	

Note: All specifications are linear regressions that include teacher experience and school fixed effects. Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school. Significance of coefficients are denoted as follows: **p < 0.01, *p<0.05, +p<0.1; p-values on tests of differences between treatment and control are shown in brackets.

Table 12: Impacts on Student Achievement Gains, School Year 2008-2009

	Math					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment School	0.024	0.024	0.022	0.040	0.057*	0.044+
	(0.017)	(0.017)	(0.019)	(0.025)	(0.024)	(0.023)
	[p=0.16]	[p=0.17]	[p=0.25]	[p=0.10]	[p=0.04]	[p=0.09]
Overall Evaluation (Pre-experiment)					0.074**	0.052**
					(0.011)	(0.011)
Value-added (Pre-experiment)						0.076**
						(0.010)
Teacher Experience Controls		√	√	√	√	√
Student-level Covariates			√	√	√	√
Restricted to Teachers in Pilot Sample				√	√	√
Sample Size	69,889	69,889	69,889	25,367	25,367	25,367

	English					
	(7)	(8)	(9)	(10)	(11)	(12)
Treatment School	-0.010	-0.012	0.011	-0.006	0.021	0.006
	(0.013)	(0.013)	(0.015)	(0.020)	(0.023)	(0.023)
	[p=0.45]	[p=0.37]	[p=0.47]	[p=0.76]	[p=0.35]	[p=0.80]
Overall Evaluation (Pre-experiment)					0.078**	0.067**
					(0.012)	(0.012)
Value-added (Pre-experiment)						0.048**
						(0.011)
Teacher Experience Controls		√	√	√	√	√
Student-level Covariates			√	√	√	√
Restricted to Teachers in Pilot Sample				√	√	√
Sample Size	67,835	67,835	67,835	23,603	23,603	23,603

Note: The dependent variables are gains in individual student test scores from 2008 to 2009, and regressions are estimated with school and teacher level random effects. P-values on tests of differences between treatment and control in brackets. *p<0.05, +p<0.1.

Table A1: Baseline Survey Responses for Treatment-Control Principals

	Control Mean	Treatment Mean	P-value H_0: T=C
Years of Experience as Evaluator	8.620	8.666	0.94
Only the Principal Contributed to the Survey	0.532	0.509	0.73
Asst. Principal also Contributed to Survey	0.404	0.474	0.30
Lead Teacher also Contributed to Survey	0.083	0.117	0.41
Other Person also Contributed to Survey	0.128	0.16	0.50
Already Monitor Test Score Growth	0.807	0.803	0.94
Top 2 Ways to Assess (Other than Observation) Include			
Student Work	0.892	0.857	0.44
State Level Standardized Tests	0.775	0.75	0.67
Feedback from Other Administrators	0.153	0.196	0.40
Feedback from Students	0.081	0.062	0.59
Teacher Work Portfolio	0.045	0.045	0.99
Feedback from Parents	0.018	0.036	0.42
Feedback from Other Teachers	0.009	0.036	0.18
Other School Related Tasks	0.009	0.018	0.57
Value Added Reports would be Extremely Useful for…			
Professional Development	0.818	0.83	0.81
Assessment of Staffing Needs	0.664	0.697	0.60
Assessment of Teachers	0.636	0.732	0.13
Assignment of Students to Teachers	0.564	0.679	0.08+
Tenure Decisions	0.545	0.607	0.35
Curricular Choices	0.436	0.526	0.18
Concerns Regarding Test Scores			
(1-5, 1 = Extremely Valid, 5 = Extremely Invalid)			
Tests Cannot Measure Other Important Outcomes	1.718	1.657	0.63
Tests do not Measure Learning Well	3.064	3.179	0.39
Tests are Biased	3.155	3.161	0.97
Teachers are Not Primarily Responsible for Test Outcomes	3.591	3.839	0.12
Tests do not Measure Our Curriculum	3.591	3.697	0.48
Level of Agreement with Following Statements			
(1-5, 1 = Strongly Agree, 5 = Strongly Disagree)			
I know who the most effective teachers are in my school	1.284	1.26	0.69
I can retain the most effective teachers in my school	1.769	1.786	0.88
I can improve my teachers' performance *(composite)*	1.884	2.000	0.17
Teachers in my school are cooperative/satisfied *(composite)*	1.927	1.944	0.81
I can select the best teachers from my applicants	2.211	2.125	0.40
I am satisfied with teaching applicants at my school	2.550	2.58	0.81
I can dismiss the least effective teachers in my school	2.789	2.893	0.54
Anyone can be an effective teacher	3.266	3.393	0.41

Note: There are 112 treatment schools and 111 control schools. P-values indicate the statistical significance of a treatment indicator to predict the survey response.

Table A2: Evidence on "Face Validity" of Treatment from Follow-up Survey

	Treatment Mean
Principal Examined Value-Added Reports	0.84
(1-5 Scale) The Value-added Reports...	
Contain Information Useful to Principals	4.29
Contain Information Useful to Teachers	4.05
Have Enhanced my Plans for Improving Instruction in my School	3.73
Have Helped Me Better Understand Differences Between Teachers	3.59
Are Easy to Understand	3.36
(1-5 Scale) How Useful Would Annual Value-Added Reports be for …	
Assigning Students to Teachers	3.89
Teacher Evaluation	3.86
Designing Professional Development for Teachers	3.76
Assessing Staffing Needs	3.59
Choices of Curricula or Instructional Programs	3.27
Confident that Value-Added Calculations Accounted for... *(Yes = 1, No = 0)*	
Teaching Experience	0.77
Prior Performance of Students on Standardized Tests	0.76
If a Teacher Recently Started Teaching a New Grade/Subject	0.53
Percentage ELL/Special Education Students in a Teacher’s Class	0.48
Which Teacher(s) the Students Had in the Previous Year	0.45
Class Size	0.40
The Number of Students who Entered the Class Mid-Year.	0.27
*Things that Distracted the Teacher's Class on the Test Day	0.18
*Outside Help a Teacher’s Students Received (e.g., after-school)	0.18
*Help a Teacher Received from an Aide in the Classroom	0.13
*Outside Help a Teacher’s Students Received (e.g., after-school)	0.10
*If a Teacher had a Personal Issue During the Year	0.08
*The Teacher's Performance in Teaching Non-tested Subjects	0.07

Note: 84 treatment schools responded to the follow-up survey, but only 79 completed the second section (after evaluating their teachers) and only 66 principals who claimed to have received and examined the reports were asked the remainder of these questions. Items marked with an asterisk (*) were not included as controls in the calculation of value-added measures.

Table A3: Baseline Evaluations and Value-Added by Subject Area

	Only Math	Only English	Math and English		
	(1)	(2)	(3)	(4)	(5)
Value-added in Math	0.385**		0.191**		0.103**
	(0.059)		(0.031)		(0.031)
Value-added in English		0.159*		0.216**	0.174**
		(0.075)		(0.030)	(0.030)
R-squared	0.51	0.38	0.32	0.33	0.34
Sample Size	544	456	1,507	1,507	1,507

Note: The dependent variable is the principal's overall evalution of teacher performance. Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Both the overall evaluation and value-added are normalized to have mean zero and standard deviaiton one for the study sample. All specifications include school fixed effects and controls for teaching experience. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.